Cross-Selling Recommendation Final Project

Virtual Internship

Kumkum Chakraborty

03/30/2025

project	Cross Selling Recommendation
Batch Code	LISUM41
Name	Kumkum Chakraborty
University	Dr.B.R.Ambedkar University
Country	U.S.A
Email	kumkumchakraborty2016@gmail.com
Specialization	Data Analyst

Agenda

• Problem Statement
• Data Information
.Data Understanding
.Exploratory Data Analysis
. Recommendations

Problem Statement:

XYZ credit union in Latin America is performing very well in selling the Banking products (e.g.: Credit card, deposit account, retirement account, safe deposit box etc) but their existing customer is not buying more than 1 product which means bank is not performing good in cross selling (Bank is not able to sell their other offerings to existing customer). XYZ Credit Union decided to approach ABC analytics to solve their problem

Objective:

This project aims to analyze customer behavior and provide **data-driven recommendations** to improve cross-selling without using machine learning.

Data Information

Column Name	Description
fecha_dato	The table is partitioned for this column
ncodpers	Customer code
ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive
pais_residencia	Customer's Country residence
sexo	Customer's sex
age	Age
fecha_alta	The date in which the customer became as the first holder of a contract in the bank

Data Information

ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.
antiguedad	Customer seniority (in months)
indrel	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
ult_fec_cli_1t	Last date as primary customer (if he isn't at the end of the month)
indrel_1mes	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
tiprel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)

Data Information

conyuemp	Spouse index. 1 if the customer is spouse of an employee
canal_entrada	channel used by the customer to join
indfall	Deceased index. N/S
tipodom	Addres type. 1, primary address
cod_prov	Province code (customer's address)
nomprov	Province name
ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
renta	Gross income of the household
segmento	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated
ind_ahor_fin_ult1	Saving Account
ind_aval_fin_ult1	Guarantees
ind_cco_fin_ult1	Current Accounts

Data Information

ind_cder_fin_ult1	Derivada Account
ind_cno_fin_ult1	Payroll Account
ind_ctju_fin_ult1	Junior Account
ind_ctma_fin_ult1	Más particular Account
ind_ctop_fin_ult1	particular Account
ind_ctpp_fin_ult1	particular Plus Account
ind_deco_fin_ult1	Short-term deposits
ind_deme_fin_ult1	Medium-term deposits
ind_dela_fin_ult1	Long-term deposits
ind_ecue_fin_ult1	e-account
ind_fond_fin_ult1	Funds
ind_hip_fin_ult1	Mortgage
ind_plan_fin_ult1	Pension

Data Information

ind_pres_fin_ult1	Loans
ind_reca_fin_ult1	Taxes
ind_tjcr_fin_ult1	Credit Card
ind_valo_fin_ult1	Securities
ind_viv_fin_ult1	Home Account
ind_nomina_ult1	Payroll
ind_nom_pens_ult1	Pensions
ind_recibo_ult1	Direct Debit

Info. About Data:

This data about XYZ credit union Company in Latin America which contains 48 features and 13647309 Number of Observations.

Total Number of Observations	13647309
Total Number of Files	1
Total Number of Features	48
Base Format of The File	CSV
Size of The Data	310MB

Data Understanding

This data set contains 13647309 rows and 48 columns

Out[2]:

	fecha_dato	ncodpers	ind_empleado	pais_residencia	sexo	age	fecha_alta	ind_nuevo	antiguedad	indrel	...	ind_hip_fin_ult1	ind_plan_fin_ult
0	2015-01-28	1375586	N	ES	H	35	2015-01-12	0	6	1	...	0	
1	2015-01-28	1050611	N	ES	V	23	2012-08-10	0	35	1	...	0	
2	2015-01-28	1050612	N	ES	V	23	2012-08-10	0	35	1	...	0	
3	2015-01-28	1050613	N	ES	H	22	2012-08-10	0	35	1	...	0	
4	2015-01-28	1050614	N	ES	V	23	2012-08-10	0	35	1	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	.

In [6]:
```
train_data.shape
```
Out[6]: (13647309, 48)

Data Understanding

All the data types was object but now few column's data types has been changed float64

In [16]:

```
print(train_data.dtypes)
```

```
record_date                        object
customer_id                        object
employee_status                    object
country_of_residence               object
gender                             object
age                               float64
customer_since                     object
new_customer_index                 object
seniority_months                  float64
primary_relationship_type          object
last_primary_relationship          object
customer_type_last_month           object
residence_flag                     object
foreigner_flag                     object
customer_acquisition_channel       object
deceased_flag                      object
address_type                       object
province_code                      object
province_name                      object
active_customer_flag               object
household_income                  float64
customer_segment                   object
savings_account                    object
```

Data Understanding

Missing Value Filling

n [25]:
```
train_data['last_primary_relationship'].fillna('Unknown', inplace=True)
train_data['customer_type_last_month'].fillna('Unknown', inplace=True)
```

n [26]:
```
import warnings
warnings.simplefilter("ignore")
```

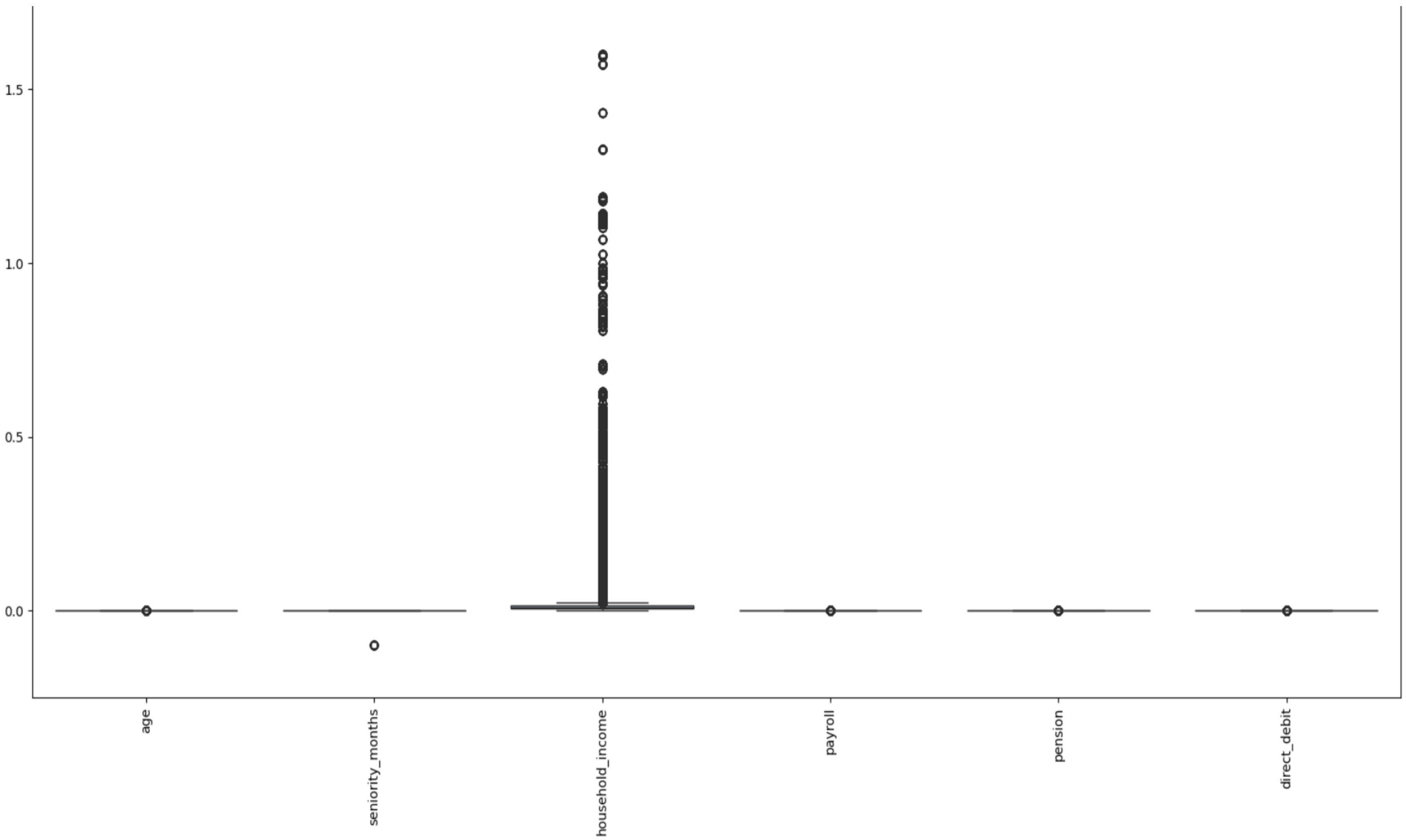
n [27]:
```
train_data['payroll'].fillna(train_data['payroll'].mode()[0], inplace=True)
train_data['pension'].fillna(train_data['pension'].mode()[0], inplace=True)
```

n [28]:
```
import warnings
warnings.simplefilter("ignore")
```

n [29]:
```
train_data['province_name'] = train_data['province_name'].fillna(method='ffill')
```

Data Understanding

Outlier Detection



Exploratory Data Analysis (EDA)

Data Describe

In [37]:

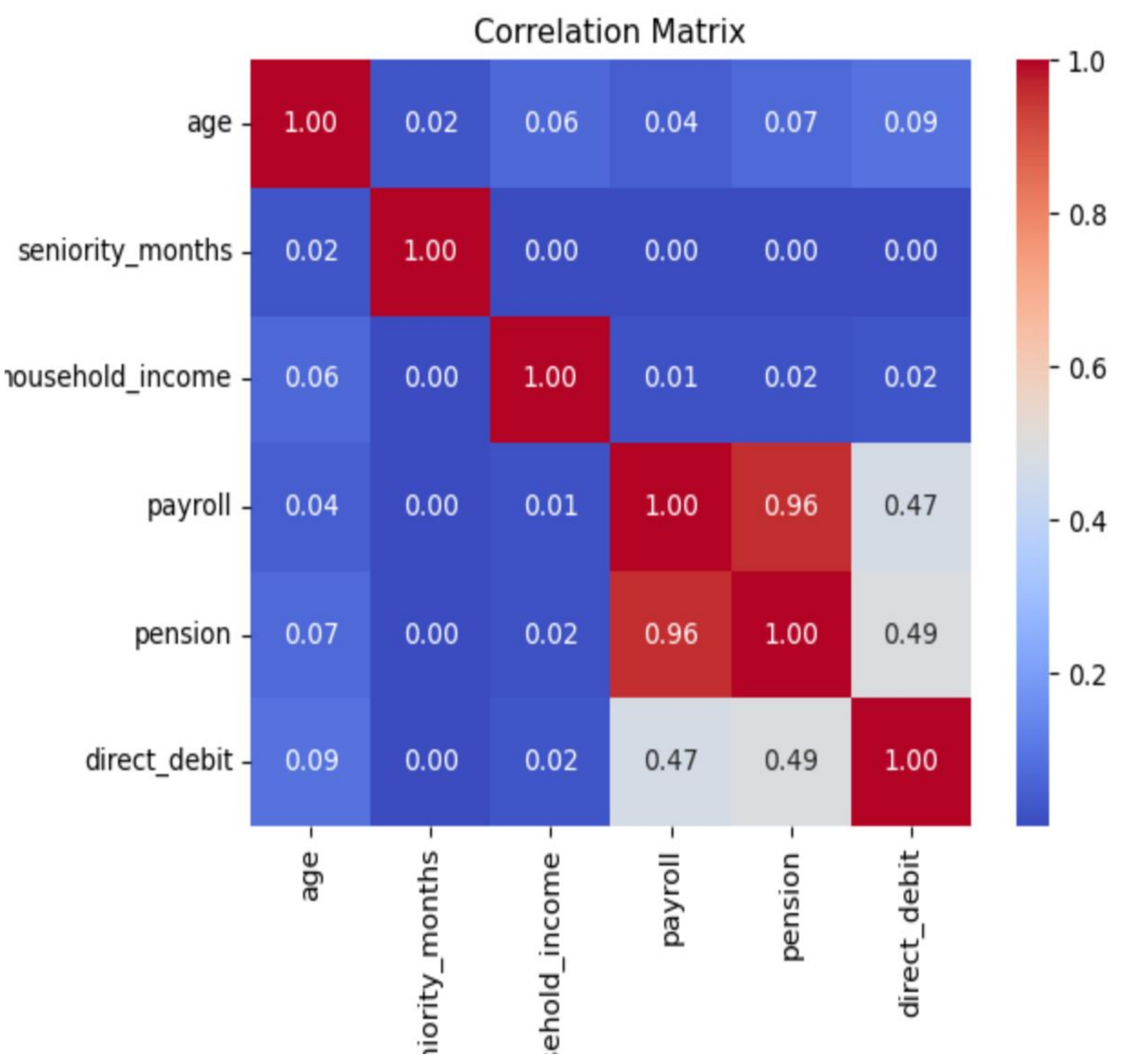
```
train_data.describe()
```

Out[37]:

	age	seniority_months	household_income	payroll	pension	direct_debit
count	1.346118e+07	1.346118e+07	1.346118e+07	1.346118e+07	1.346118e+07	1.346118e+07
mean	4.024752e+01	7.733650e+01	1.278499e+05	5.536326e-02	6.012495e-02	1.295556e-01
std	1.715972e+01	1.681596e+03	2.071576e+05	2.286879e-01	2.377182e-01	3.358139e-01
min	2.000000e+00	-9.999990e+05	1.202730e+03	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.400000e+01	2.300000e+01	7.610586e+04	0.000000e+00	0.000000e+00	0.000000e+00
50%	3.900000e+01	5.100000e+01	1.018500e+05	0.000000e+00	0.000000e+00	0.000000e+00
75%	5.000000e+01	1.360000e+02	1.381542e+05	0.000000e+00	0.000000e+00	0.000000e+00
max	1.640000e+02	2.560000e+02	2.889440e+07	1.000000e+00	1.000000e+00	1.000000e+00

Exploratory Data Analysis (EDA)

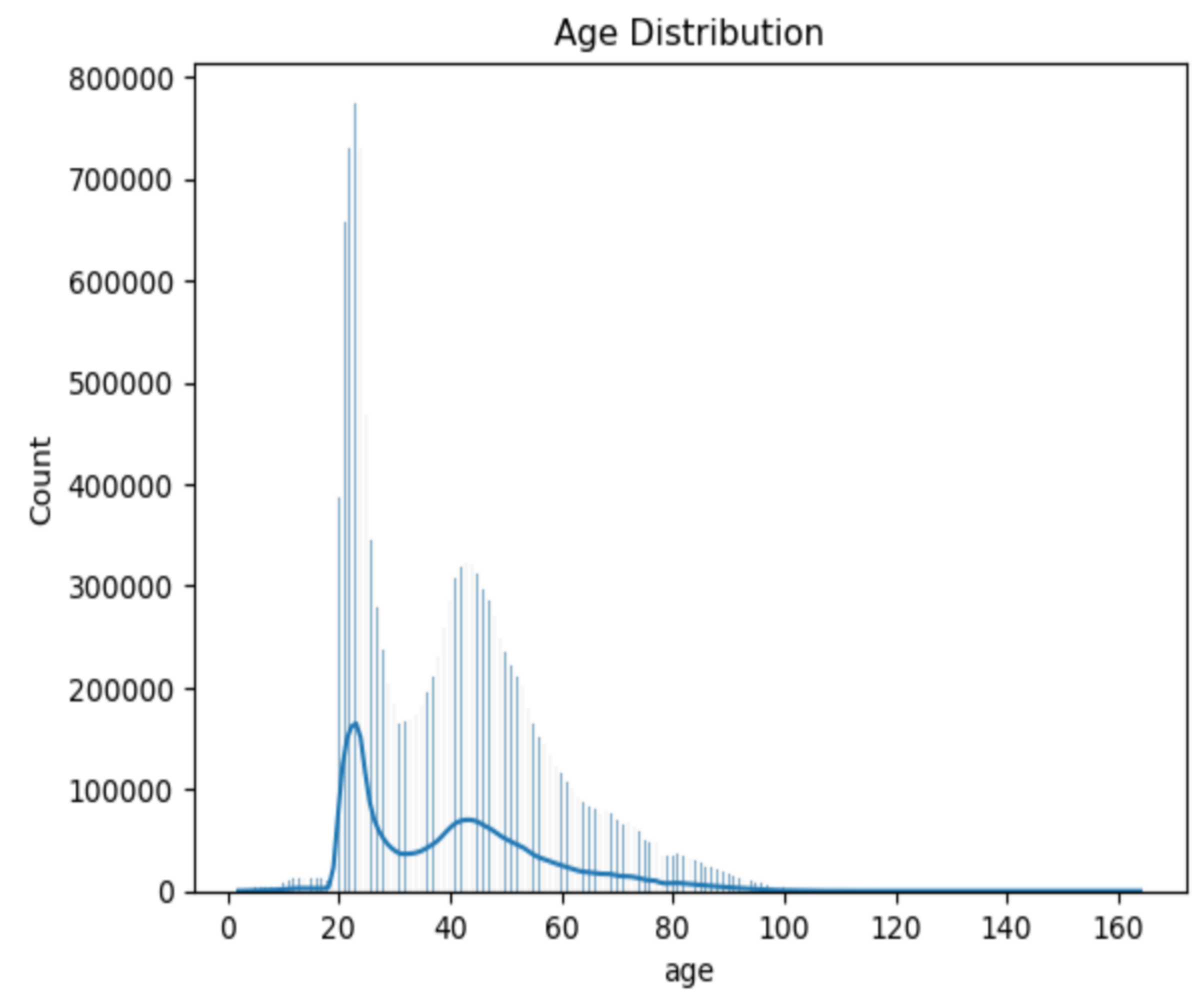
A strong correlation (0.96) between payroll and pension accounts suggests a high cross-selling opportunity, enabling the bank to target payroll account holders for pension products



Exploratory Data Analysis (EDA)

1.The majority of customers are between 19 to 25 years old, as shown by the highest frequency in this range.

2. The age distribution is right-skewed, indicating fewer older customers in the dataset

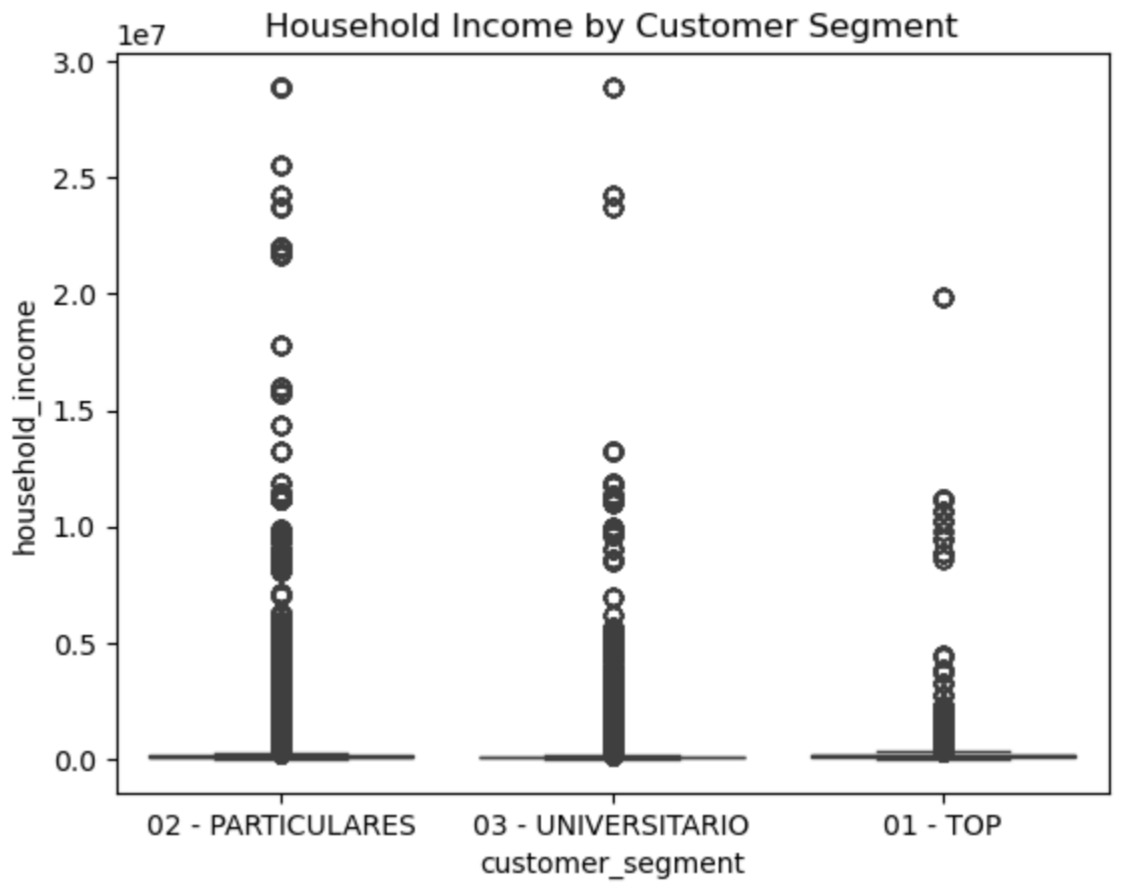


Exploratory Data Analysis (EDA)

1.The **02 (Particulars)** segment has the highest concentration of household income up to **1.0**, with some high-income outliers reaching **3.0**

2. The **03 (Universitario)** segment has most customers with household income below **0.6**, but a few outliers extend beyond **3.0**

3. The **01 (Top)** segment has lower household income distribution, mainly below **0.6**, with very few high-income customers reaching **2.0**



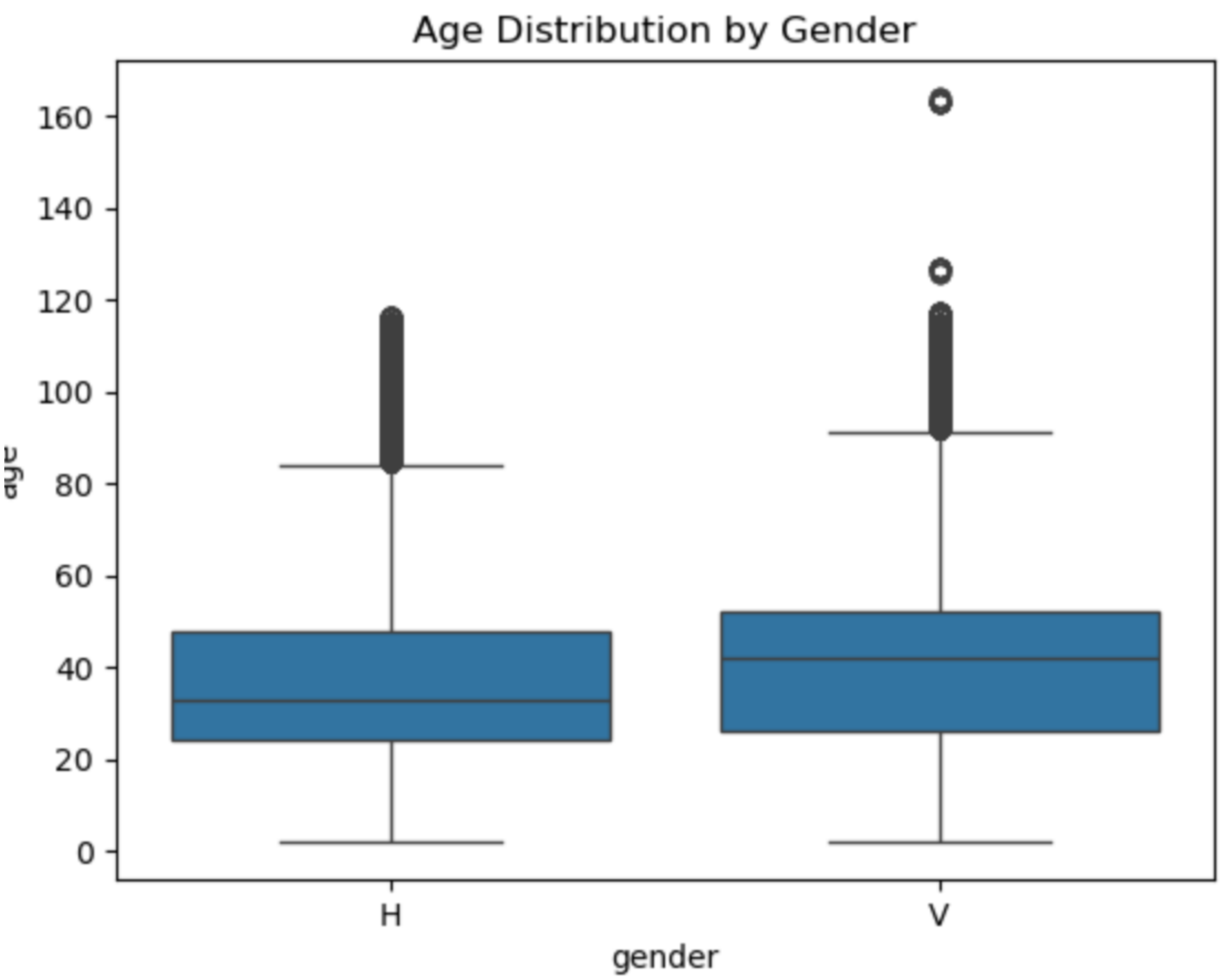
This suggests that the **VIP segment has fewer high-income customers**, while the **Particulars and Universitario segments show a wider spread of income levels**, which could impact cross-selling strategies

Exploratory Data Analysis (EDA)

- 1. The median age for gender 'H' is around 30, while for gender 'V,' it is around 40

- 2. Both genders have a similar upper range (90-120), but gender 'V' has more extreme outliers at ages 120 and 160.

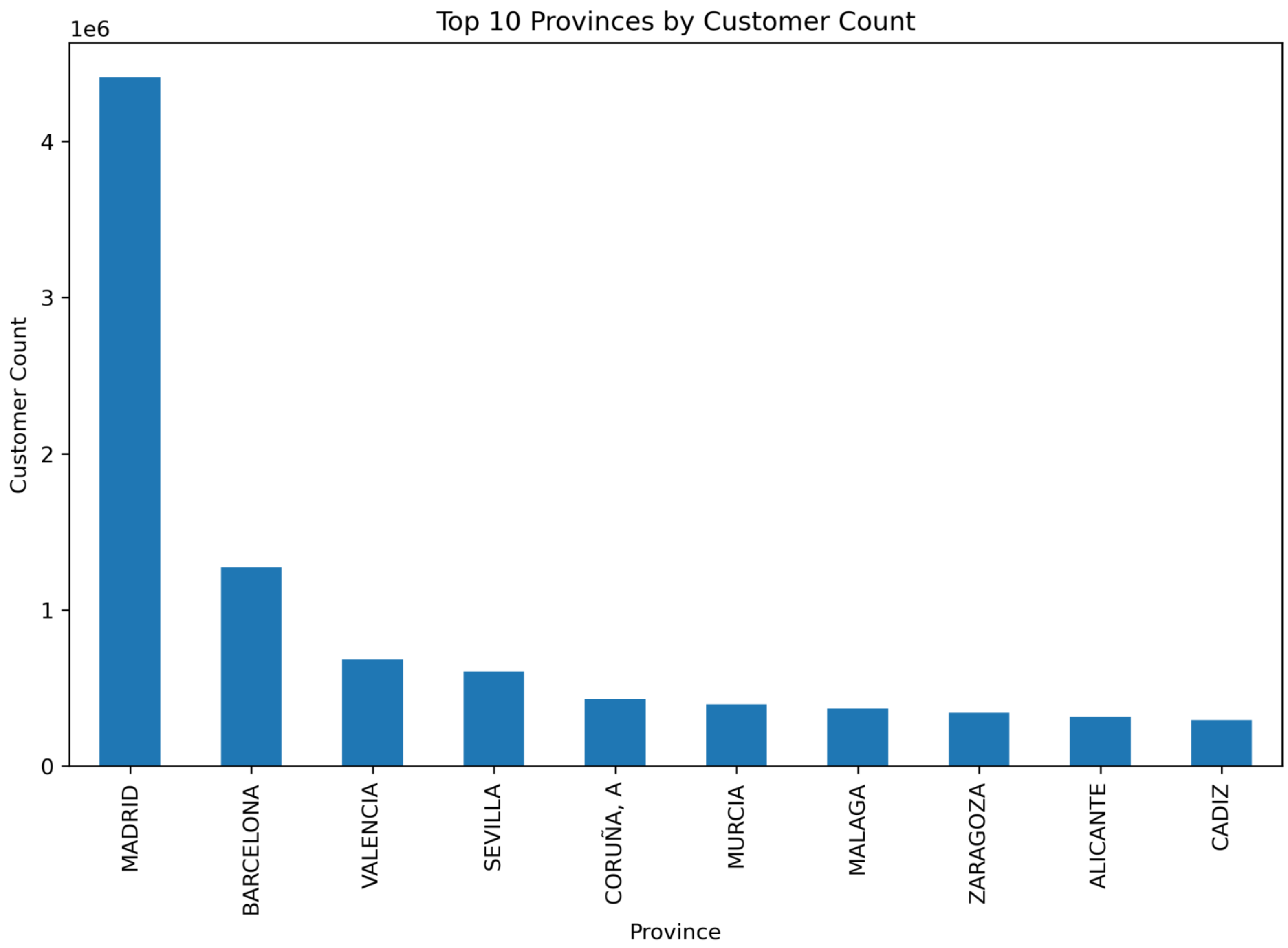
- 3. Gender 'H' has a more compact distribution, whereas gender 'V' shows a wider spread in age



Exploratory Data Analysis (EDA)

1.Madrid has the highest number of customers, with its count exceeding 400,000, significantly higher than other provinces

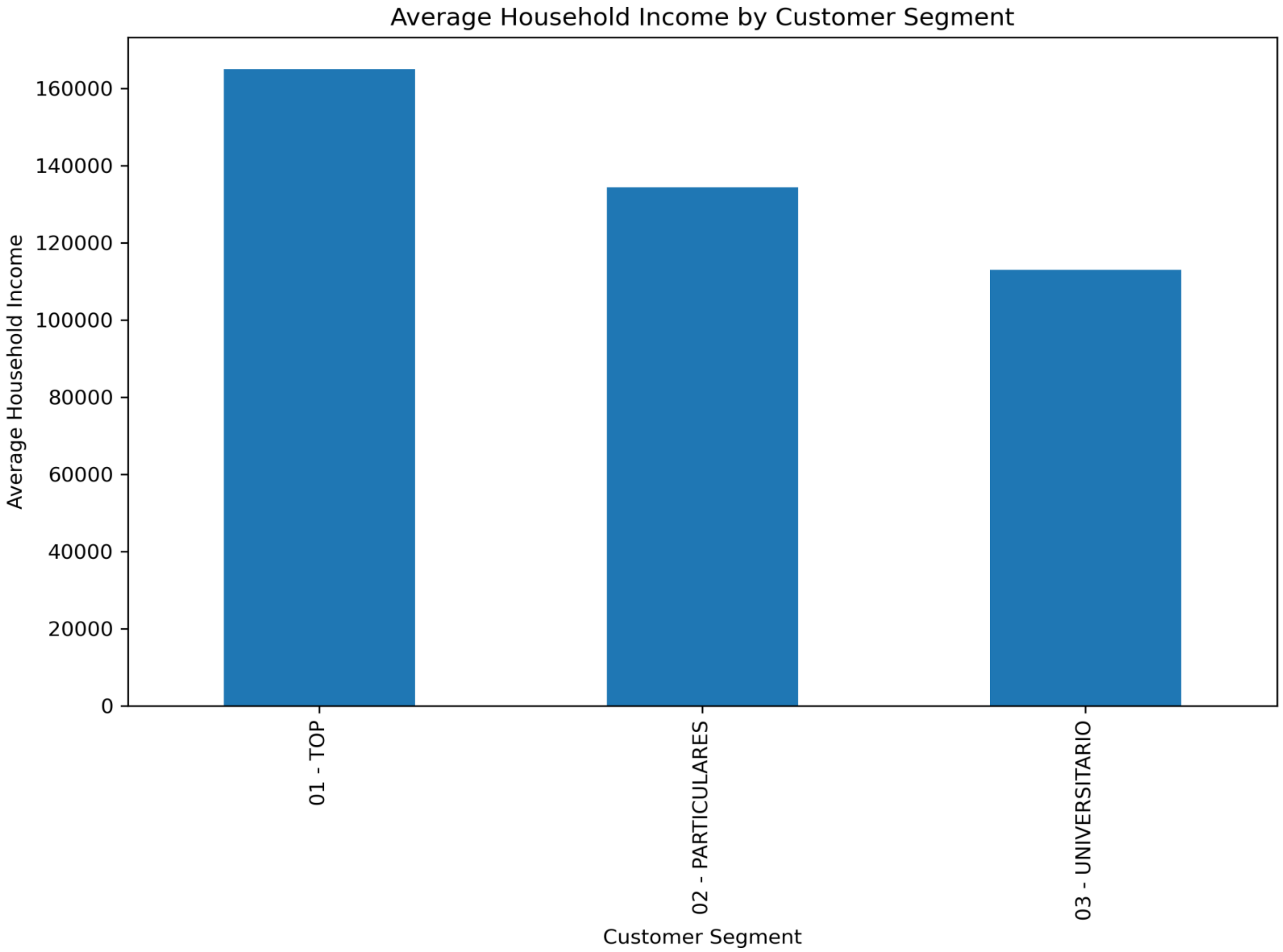
2. Other provinces have less than 200,000 customers, indicating that Madrid is the primary market for the bank's products



Exploratory Data Analysis (EDA)

1.The top (01) segment has the highest average household income, reaching **160,000**

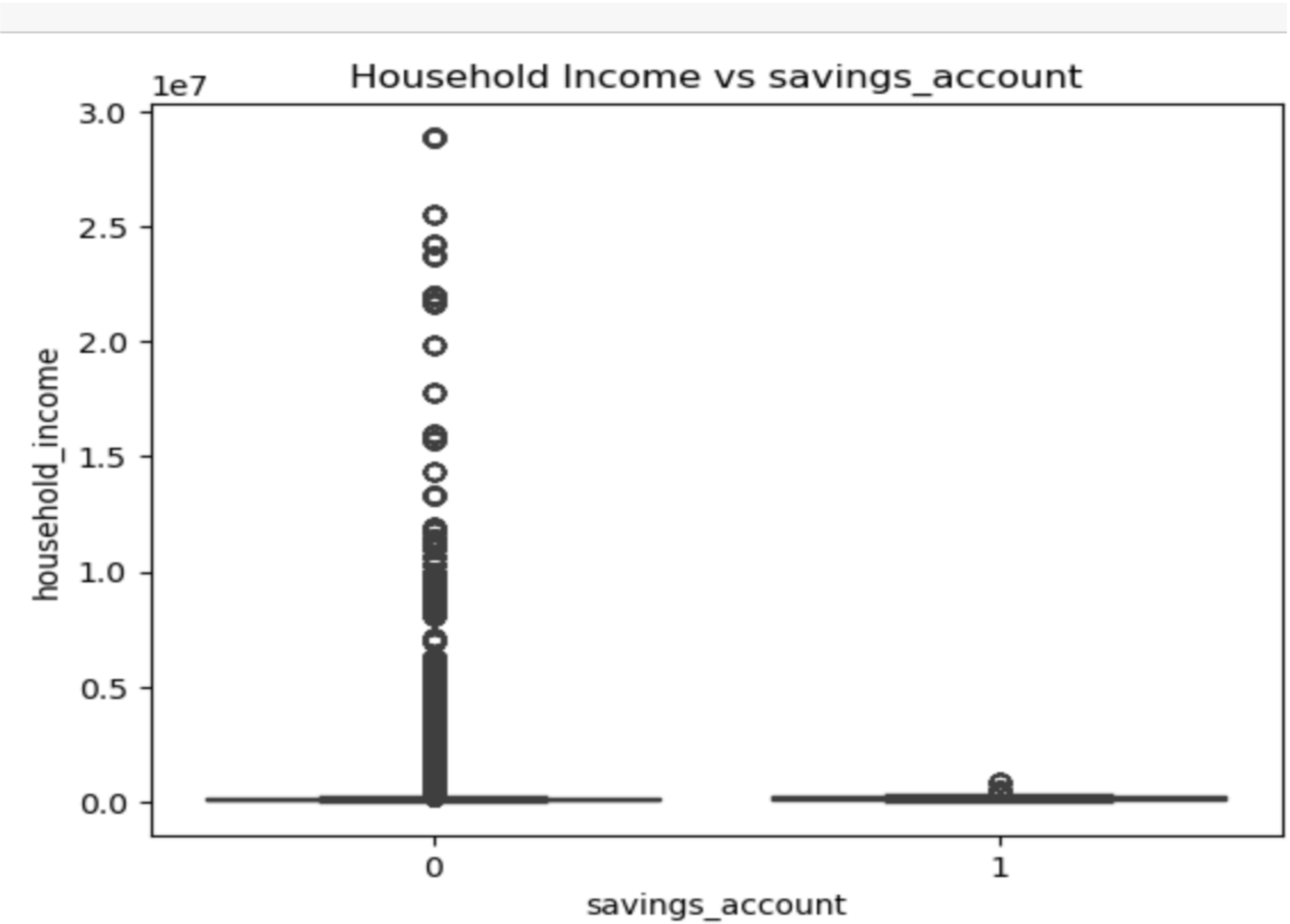
2. The **Particular (02)** segment follows with **120,000**, while the **University (03)** segment has the lowest at **100,000**



Exploratory Data Analysis (EDA)

1.Customers without a savings account (0) show a wide income distribution, with most data points concentrated below **1.3le7** and a few outliers reaching **3.0le7**

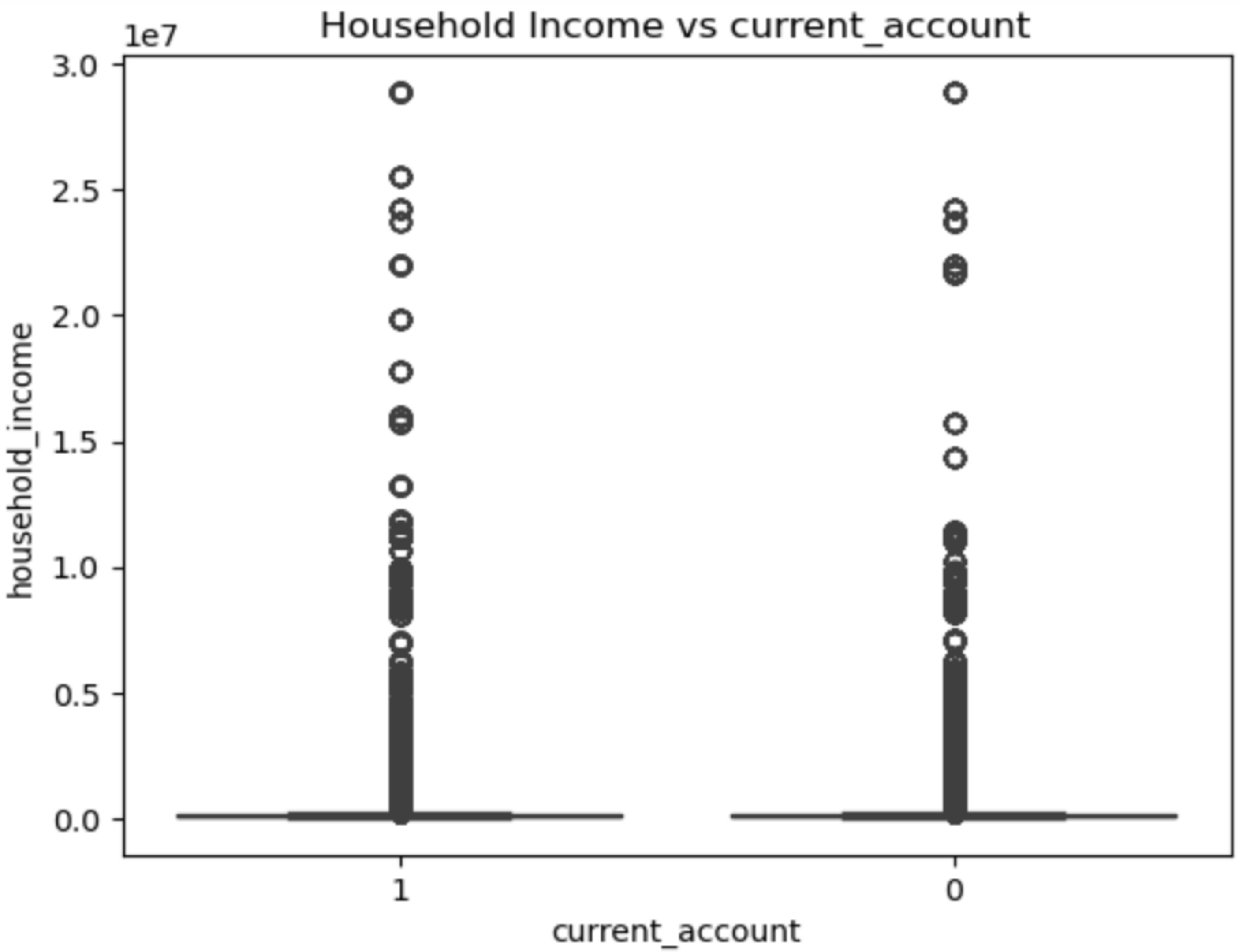
2. Customers with a savings account (1) have significantly lower income distribution, with most values clustered below **0.2le7**, indicating a trend where higher-income individuals may not prioritize savings accounts.



Exploratory Data Analysis (EDA)

1.Customers **without** a current account (0) have a wide income distribution, with most values **densely concentrated below 1.3 le7**, while a few outliers reach **3.0 le7**

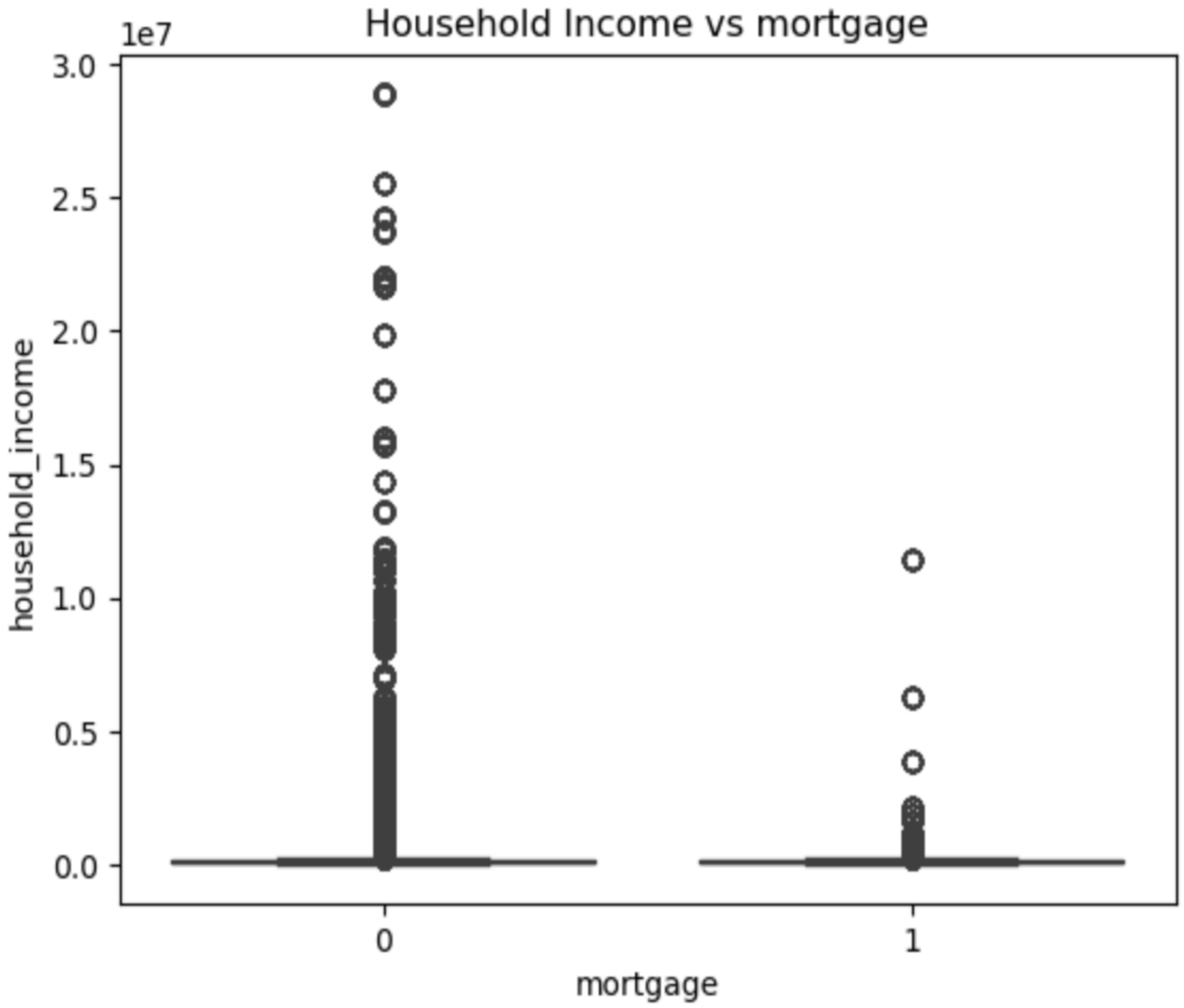
2. Customers **with** a current account (1) also show a similar pattern, with most incomes **below 1.2le7**, but outliers extend up to **3.0le7**, suggesting that income level alone may not strongly determine current account ownership



Exploratory Data Analysis (EDA)

•**Customers without a mortgage (0)** have household income **densely concentrated below 1.3 le7**, with scattered high-income outliers reaching **3.0 le7**.

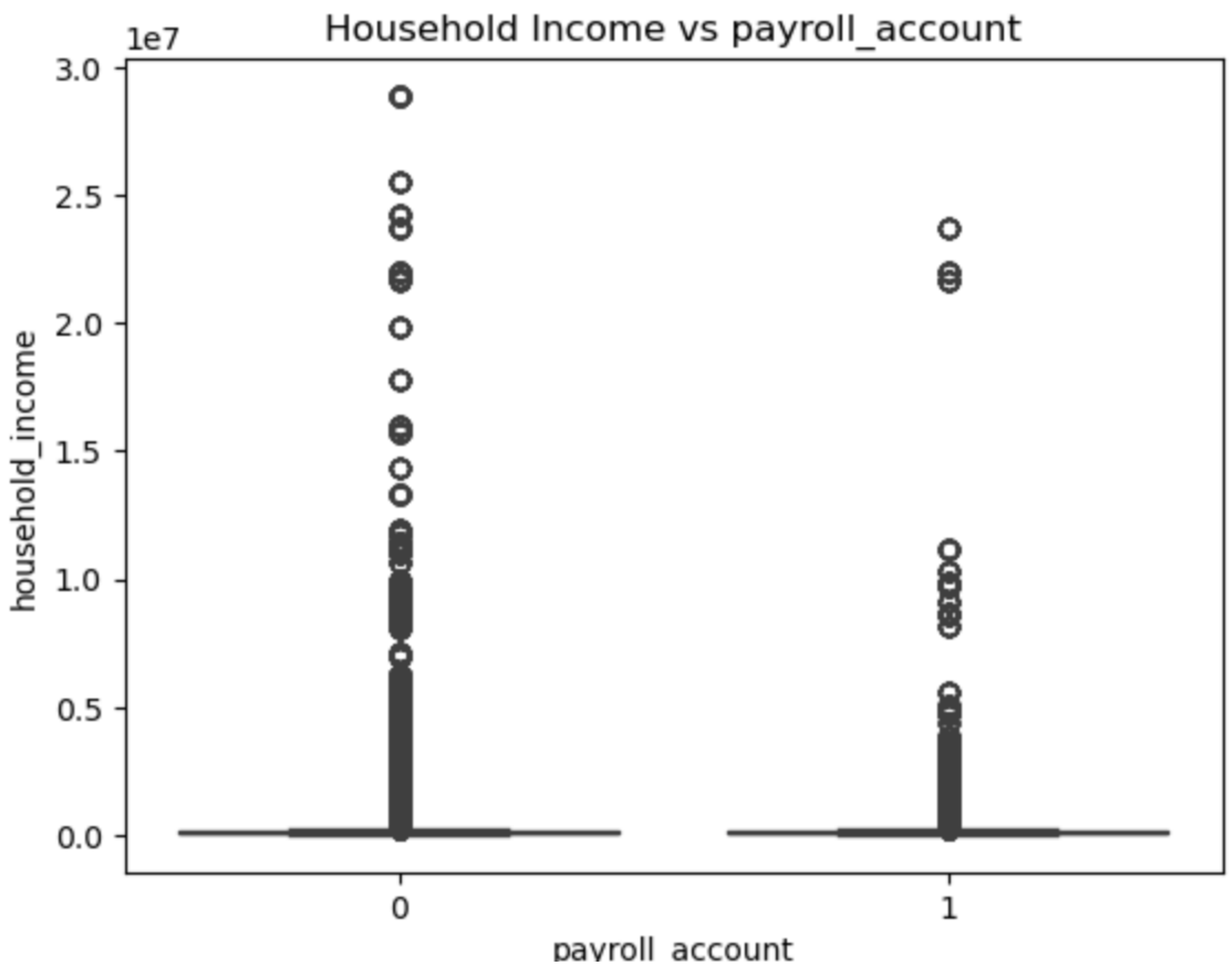
Customers with a mortgage (1) have household income **densely concentrated below 0.3 le7**, with a few scattered high-income cases up to **1.2 le7**



Exploratory Data Analysis (EDA)

1.Customers without a payroll account (0) have household income **densely concentrated below 1.3 le7**, with scattered high-income outliers reaching **3.0 le7**.

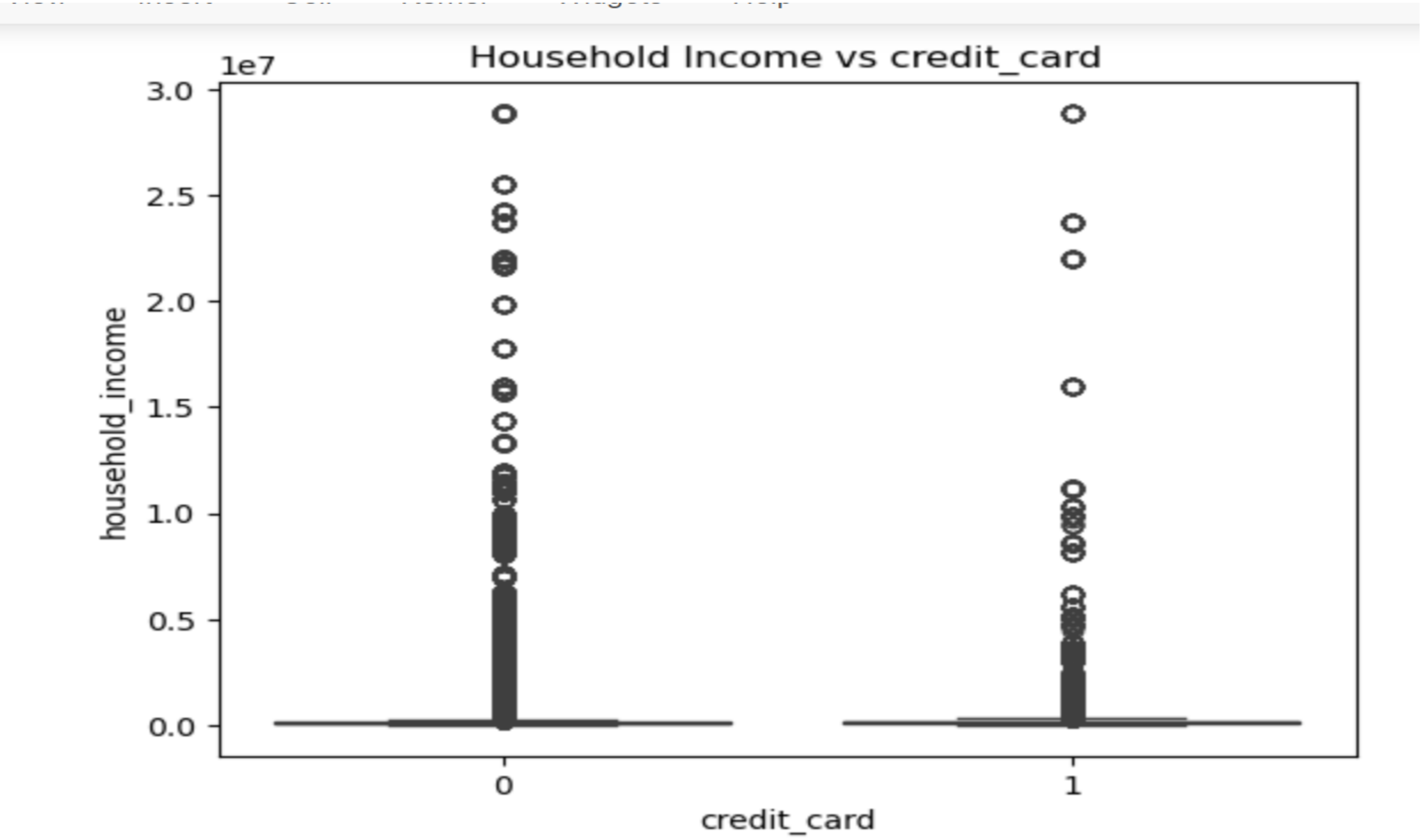
Customers with a payroll account (1) have household income **densely concentrated below 0.6 le7**



Exploratory Data Analysis (EDA)

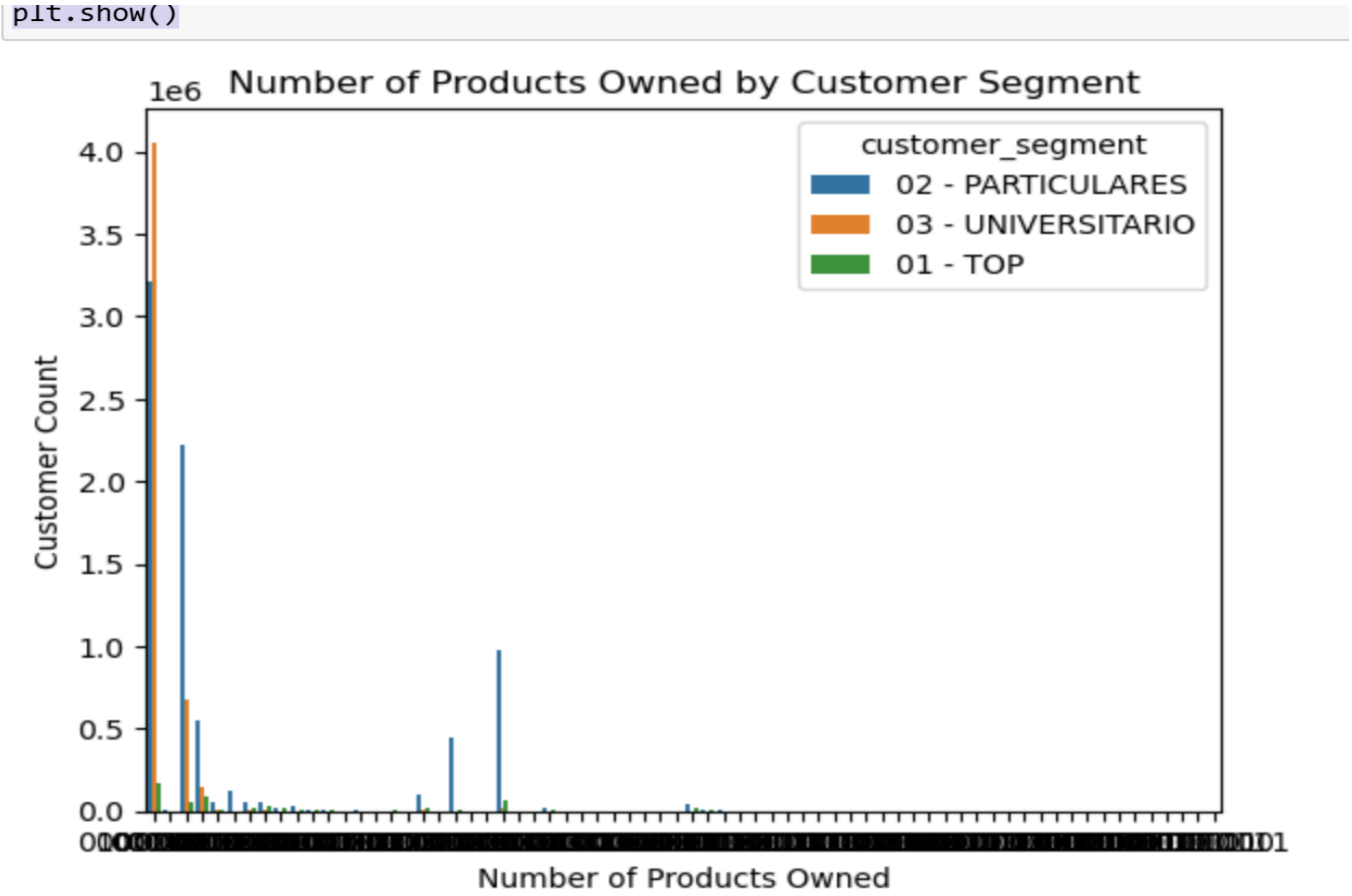
Household income is **densely concentrated below 1.2 le7**, with scattered high-income outliers reachi **3.0 le7**

Household income is **densely concentrated below 0.5 le7**, with some overlapping values till **1.2 le7**, and a few high-income cases up to **3.0 le7**.



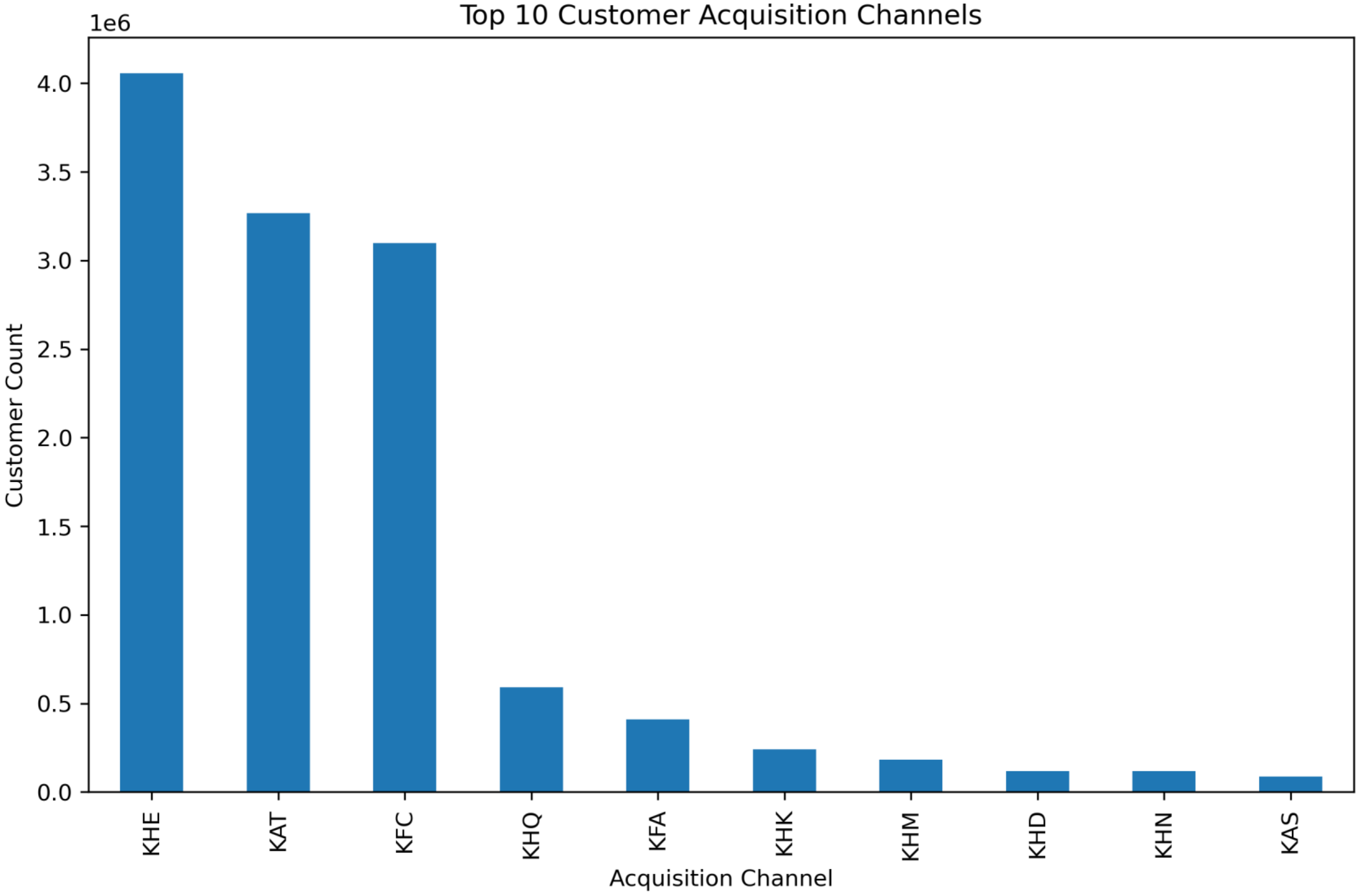
Exploratory Data Analysis (EDA)

Customers owning **four** products are predominantly from the **Universitario**, indicating a strong preference for multiple banking products in this group



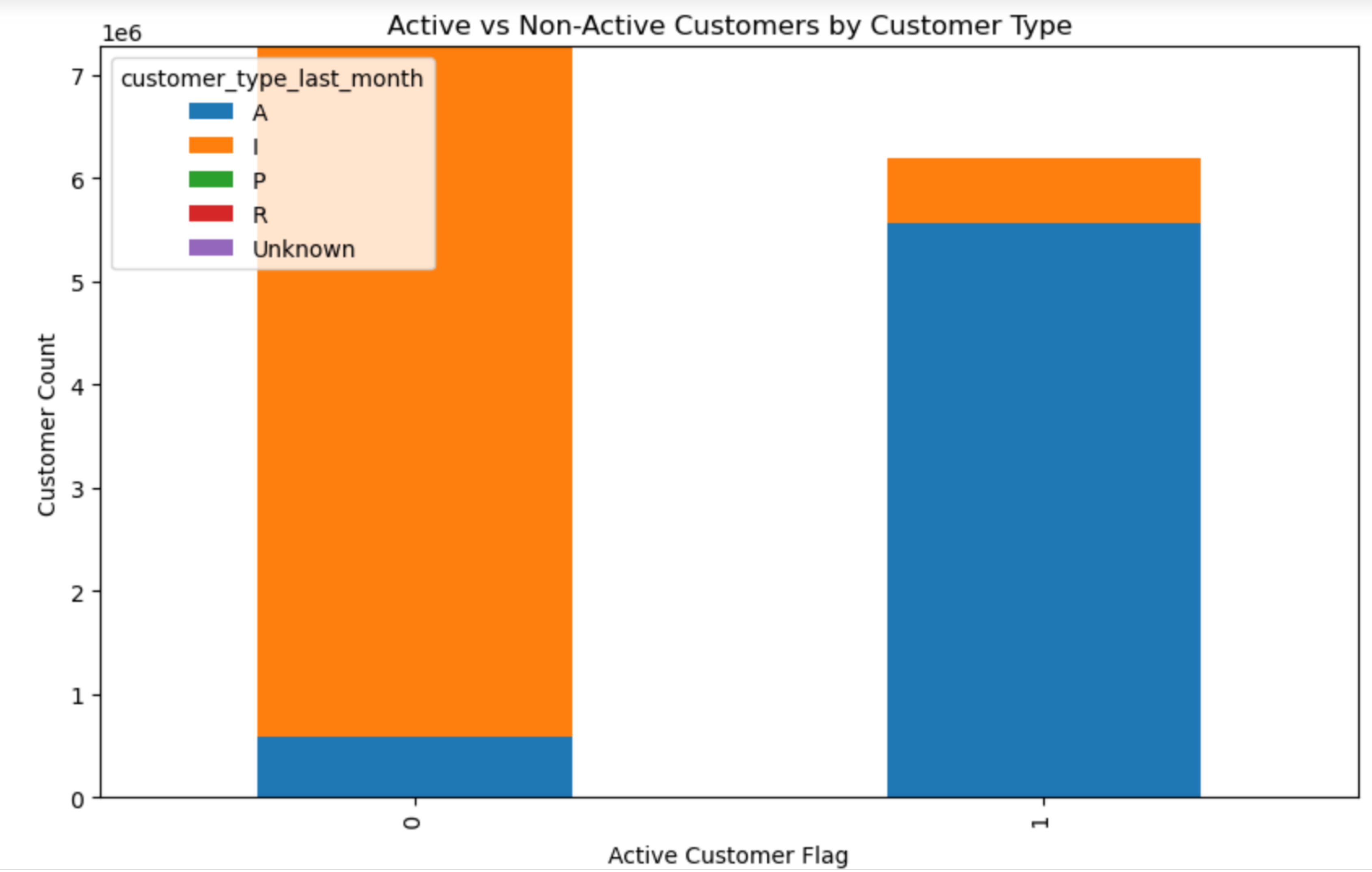
Exploratory Data Analysis (EDA)

Customer acquisition is primarily driven by a few key channels, with 'KHE' being the most dominant, bringing in around 4.0 le6 customers. Other significant channels include 'KAT' (3.5le6) and 'KFC' (3.3le6)



Exploratory Data Analysis (EDA)

We see active customer flag is A and few I.

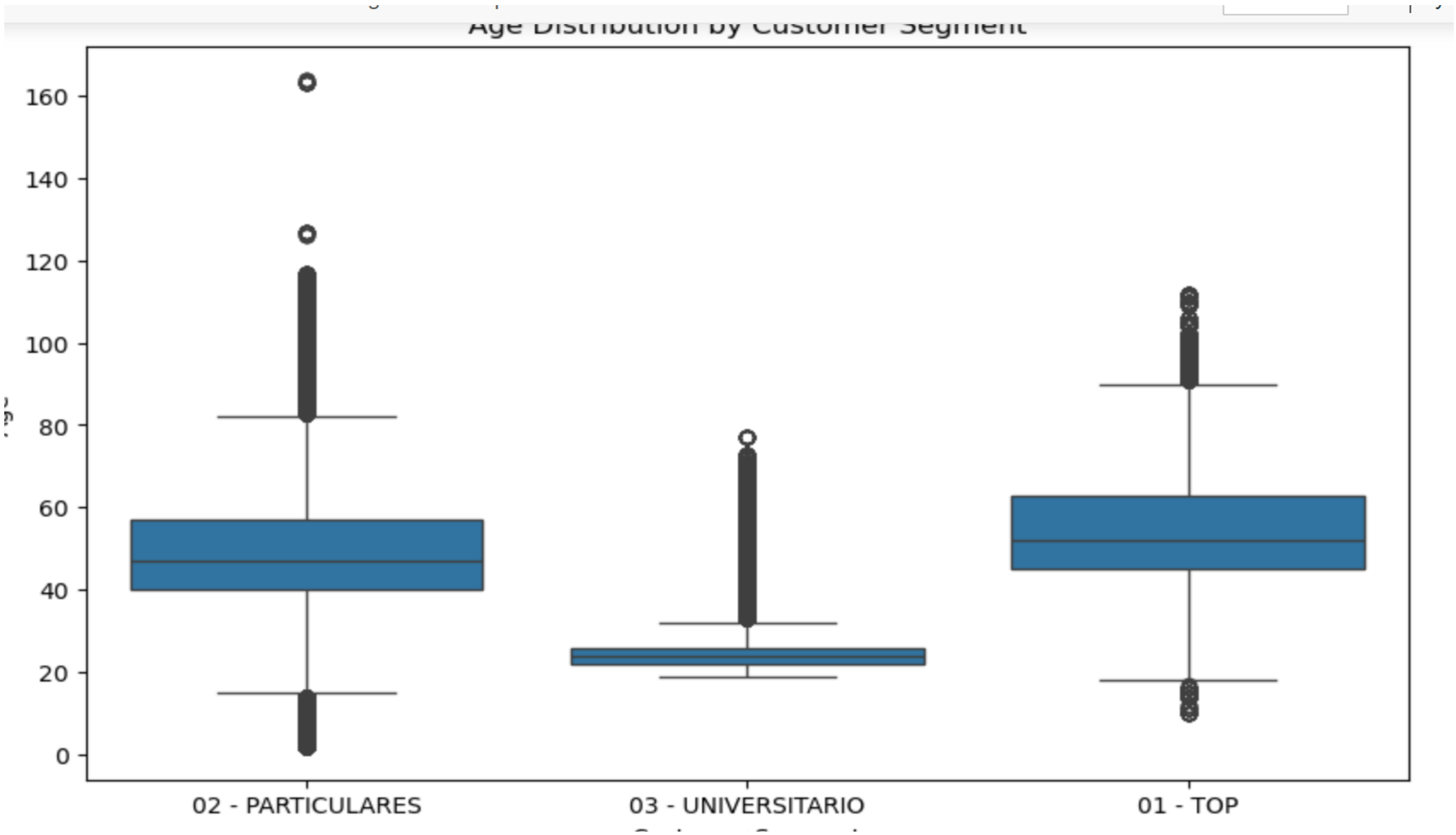


Exploratory Data Analysis (EDA)

Age Distribution by Customer

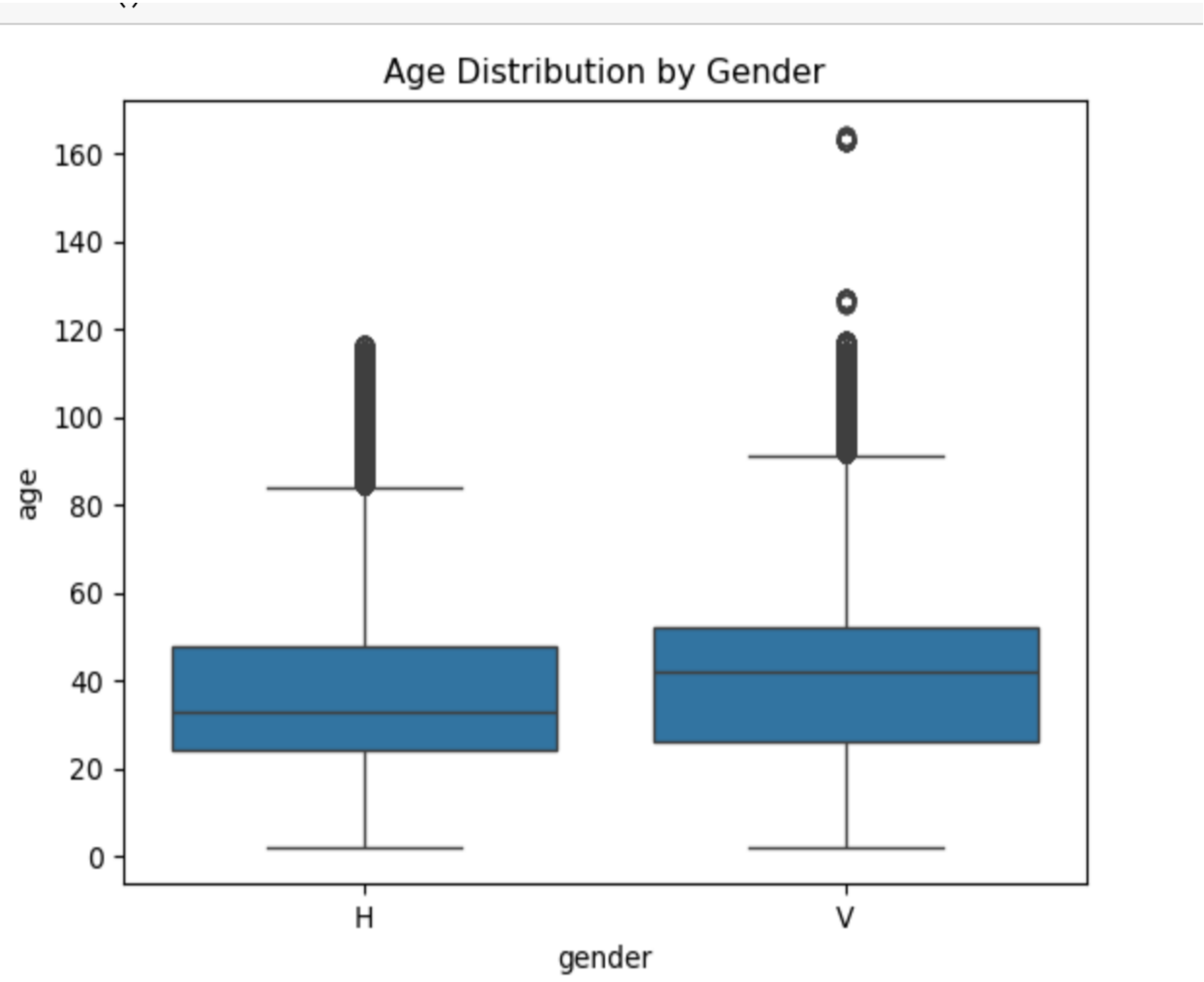
1.Customers in **Segment 01 (Top)** and **Segment 02 (Particulars)** have a similar age distribution, with most customers around **45 years old**, but some older customers (82-120 years) are present.

2.Segment 03 (University students) has a younger age group, with most customers around **25 years old**, and very few customers above **75 years**



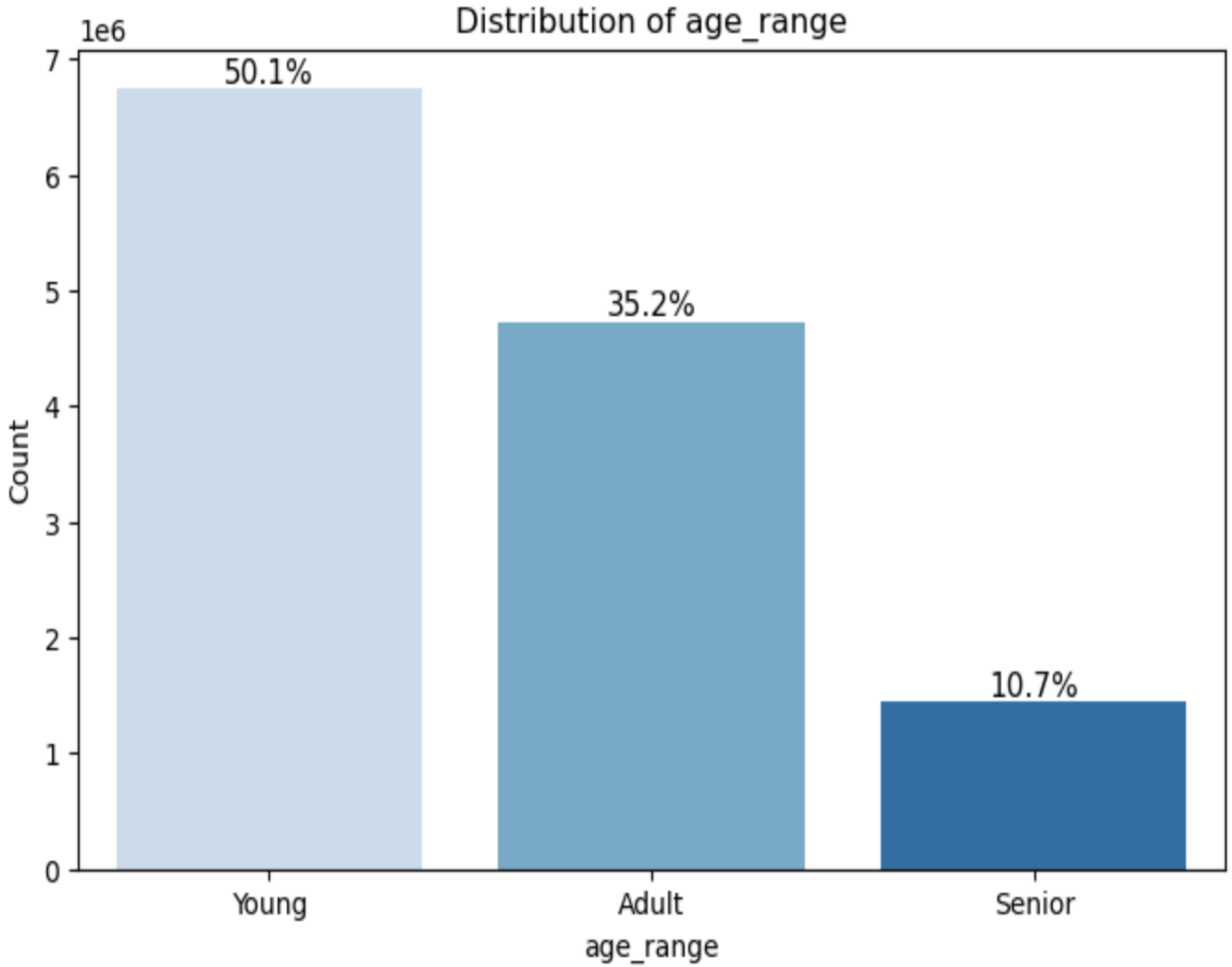
Exploratory Data Analysis (EDA)

From this graph, we see gender **‘V’** Aage is higher than **H**



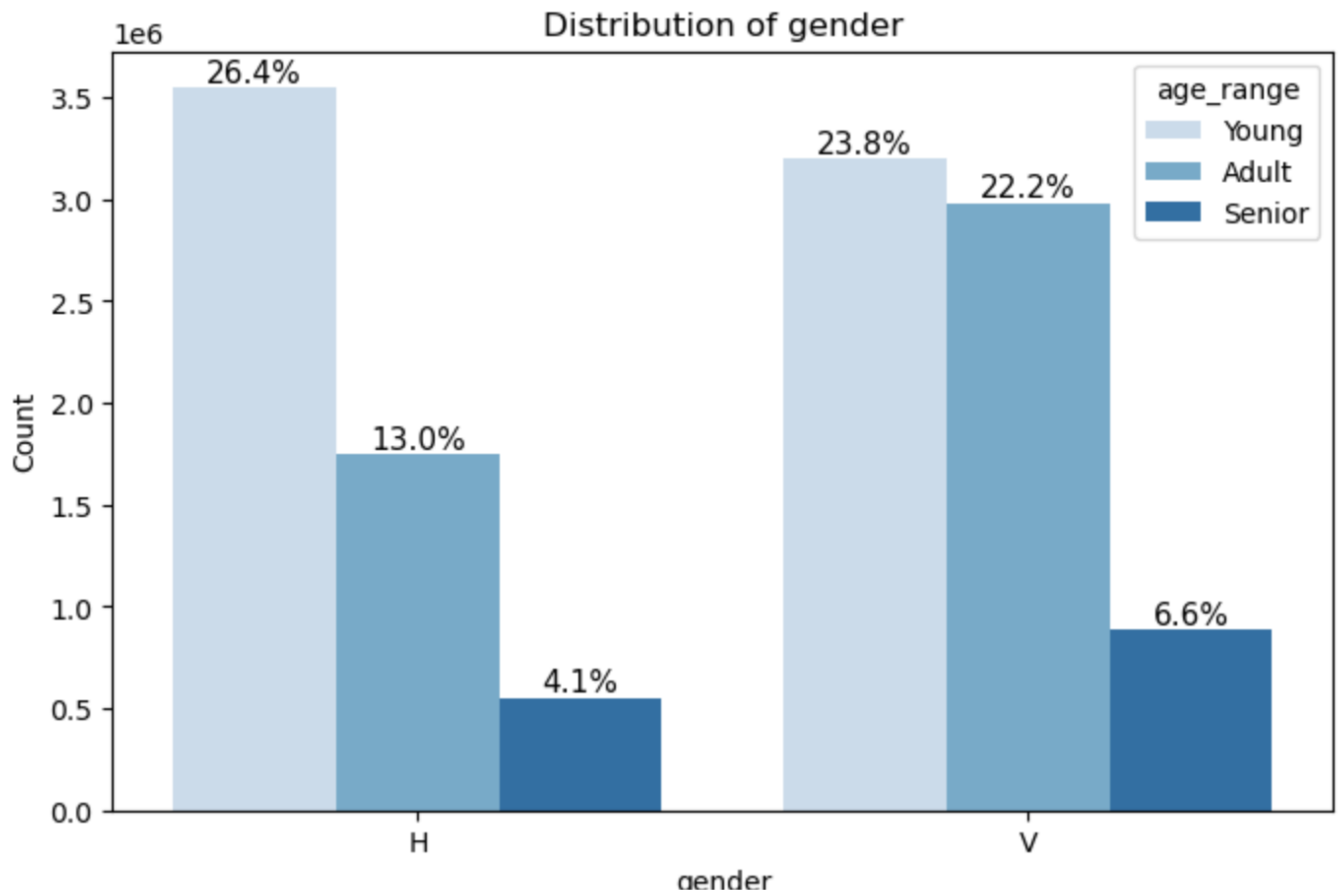
Exploratory Data Analysis (EDA)

Young age is more than half and then adult.



Young customers are slightly more **H (26.4%)** than **V (23.0%)**

```
In [62]: plot(feature='gender',data=train_data, hue='age_range')
```



Final Recommendations

1. Prioritize High-Income Customers (Segment 01) for Premium Products

1. Provide Incentives for Low-Income Customers (Segment 03) to Adopt More Products

1. Offer low income customers with exclusive credit cards, investment accounts, and premium banking services

1. Focus Marketing Efforts on Madrid (Largest Customer Base)

1. Increase branch promotions and personalized marketing in this Madrid

2. Target Young Customers (19-25) with Digital Banking & Student Loans, Promote mobile banking, student credit cards, and small personal loans for them

Thank you